To bind or not to bind

Mark D. Biggin

Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. e-mail: mdbiggin@lbl.gov

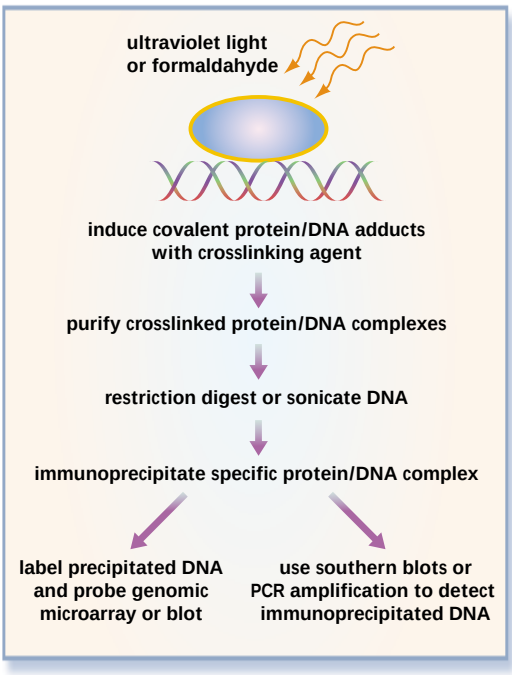
Gene expression is regulated by transcription factors binding selectively to particular portions of the genome. To what extent are these protein–DNA interactions influenced by the intrinsic sequence-specific recognition properties at each protein, and to what extent are they affected by other factors, such as chromatin structure or cooperative interactions with other proteins. Genome-wide surveys of DNA binding by transcription factors *in vivo* are beginning to provide some answers.

As the immunologist Paul Ehrlich emphasized, almost all biochemical reactions require the specific interaction of a protein with one or more ligands, be they small molecules, nucleic acids or other proteins. Measurements *in vitro* of the affinity and specificity of protein–ligand interactions have long been available for many classes of protein. These interactions are, however, almost certainly modified by the complex intracellular milieu of competing ligands and other molecules. If we could measure specific protein–ligand interactions *in vivo*, not only would we obtain a more accurate description of how proteins behave; but also, by comparing binding *in vitro* and *in vivo*, we could begin to understand how the intrinsic binding specificities of proteins are modified in cells.

> *Corpora non agunt nisi ligata.*
> (A substance is not effective unless it is linked to another.)
> —Paul Ehrlich

But quantitating interactions *in vivo* has proven much more difficult than it has *in vitro*, and so there is no *in vivo* binding data for most classes of protein. Over the past ten years, however, methods of covalently crosslinking proteins to DNA *in vivo* have been developed that can measure the interaction of sequence-specific DNA-binding proteins with genomic DNA. The earliest work was limited to studying the interactions of proteins with just a handful of gene fragments[2–8] (Fig. 1). More recent experiments[9,10], including those of Jason Lieb and colleagues reported in this issue (see page 327)[11], have combined DNA microarray technology with *in vivo* crosslinking to measure the interaction of several sequence-specific transcription factors in yeast with thousands of potential binding sites throughout the yeast genome (Fig. 1).

Sequence-specific DNA-binding proteins from eukaryotes generally recognize degenerate motifs of 5–10 base pairs. Consequently, potential recognition sequences for many transcription factors occur frequently throughout the genome. It is generally assumed that many of these potential recognition sites are not occupied to any significant extent; and in support of this, DNA binding at a few sites has been shown to be either inhibited by chromatin structure or increased by cooperative interactions with other proteins[12–14]. But without information on a sufficiently large sample of potential binding sites, it has not previously been possible to make statistically valid comparisons of the differences between *in vitro* and *in vivo* DNA binding patterns. The recent genome-wide surveys in the yeast *Saccharomyces cerevisiae*[9–11] provide the first large-scale comparisons of *in vivo* protein–DNA crosslinking to the frequency of probable recognition motifs in genomic DNA.



Bob Crimi

Fig. 1 Outline of *in vivo* crosslinking methods. Either ultraviolet (UV) light or formaldehyde is used to covalently couple endogenous proteins to their target DNAs in living cells. After immunoprecipitation of the purified protein/DNA complexes, early experiments identified the bound DNA either by labeling the immunoprecipitated DNA and using this to probe blots containing cloned DNA fragments, or by PCR amplifying potential target DNAs using specific primers, or by assaying the immunoprecipitated DNA on southern blots probed with cloned gene fragments [2–8]. The new microarray assay labels the immunoprecipitated DNA and uses it to probe a genomic microarray that contains several thousand genomic DNA fragments [9–11].

A detailed comparison has been made for two transcription factors—SBF and Rap1—which crosslink *in vivo* to between 2–5% of the fragments of yeast genomic DNA present on the microarrays. What is the outcome? *In vivo*, both SBF and Rap1 preferentially occupy their potential recognition sites in promoter sequences; sites in protein-coding regions and in non-promoter intergenic regions are occupied much less frequently. For example, of non-telomeric regions that contain probable Rap1-recognition motifs, 182 out of 322 (57%) intergenic regions were bound *in vivo*, whereas only 23 out of 163 (14%) protein-coding regions were scored as binding. Among intergenic DNA regions containing probable Rap1 recognition motifs, 46% of presumed promoter regions were bound, but only 17% of non-promoter sequences were apparently occupied. As Lieb *et al.* note[11], the numbers of DNA fragments determined as bound or not bound *in vivo* in these studies are probably not entirely accurate. There are several reasons for this, including some inexactness in the current crosslinking/microarray assay and

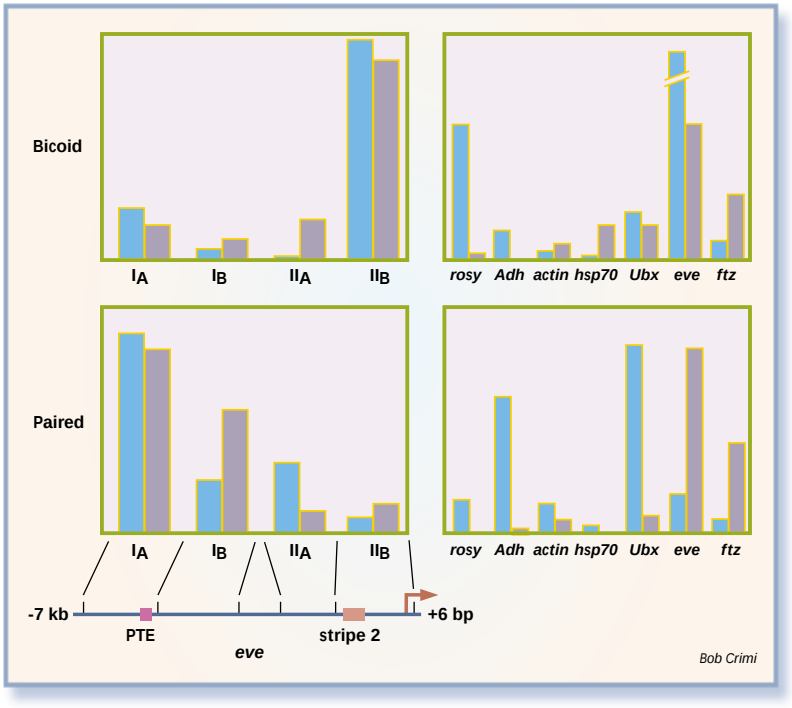


Fig. 2 Comparison of relative *in vitro* (blue) and *in vivo* (gray) DNA binding of two *Drosophila* transcription factors to the same DNA fragments[8]. The y axis is an arbitrarily scaled, relative measure of DNA binding per kb of DNA. ***a***, *In vitro* binding of Bicoid and Paired to four DNA fragments of the *eve* promoter is broadly similar to the pattern of binding *in vivo*. At the bottom is a diagram of the *eve* gene, showing the Paired target element (PTE), the Bicoid-responsive stripe 2 element (stripe 2), and the mRNA start site (arrow). ***b***, *In vitro* and *in vivo* DNA binding preferences of these two transcription factors differ when compared across genes that are transcribed to different extents. The *eve* and *ftz* genes are strongly transcribed, whereas the other genes are either not transcribed or are only weakly transcribed in a subset of cells.

in our understanding of the location of coding and non-coding DNA in the yeast genome. The trend of the data is clear, however: conditions in the cell modify the effective DNA-binding specificities of two unrelated transcription factors in a similar manner, implying that a general mechanism may be responsible.

There are two ways in which the intrinsic DNA-binding specificity of a protein can be modified *in vivo*. DNA binding can be either inhibited at certain sites—for example, by chromatin structure rendering the site inaccessible to the transcription factor—or selectively increased at certain sites by cooperative association with other proteins. In a system using only inhibition, a specific pattern of binding can only be achieved if the transcription factor is expressed at a sufficient concentration to significantly occupy those sites in the DNA at which binding is not blocked. In contrast, in a system using only selective enhancement, the transcription factor must be at a sufficiently low concentration that it cannot occupy its recognition sites at functionally significant levels without cooperative associations. At present, we do not know which of these two mechanisms is responsible for modifying the binding of Rap1 and SBF to DNA *in vivo*, or whether a combination of the two mechanisms pertains. Related experiments in *Drosophila* do show, however, that chromatin structure inhibits the binding of some transcription factors *in vivo*, suggesting that this mechanism could account, at least in part, for the yeast data.

The *Drosophila* experiments also compared *in vivo* crosslinking and *in vitro* DNA binding, but the *in vitro* data were derived by directly measuring the affinity of transcription factors for DNA fragments, and binding was examined for only a few genes[8]. A sample of the results for two of the five transcription factors studied is given in Fig. 2. The two factors—Paired and Bicoid—bind similarly *in vivo* and *in vitro* across highly transcribed genes, as illustrated for their binding to the *eve* gene. But when genes that are transcribed to different extents are examined, no simple correlation between *in vitro* and *in vivo* binding is observed. One interpretation of these results is that for highly transcribed target genes such as *eve* and *ftz*, binding sites in regulatory DNA are fully accessible, and Paired and Bicoid are expressed at sufficiently high levels to occupy these sites—hence the similarity between their *in vitro* and *in vivo* binding preferences across these genes. For weakly transcribed genes, on the other hand, binding sites might not be fully accessible as a result of the structure of the chromatin[13,14]. Binding of transcription factors would therefore be greatly reduced, explaining why genes such as *rosy* and *Adh* are bound *in vivo* more weakly by Paired and Bicoid than expected from the intrinsic *in vitro* affinity of these proteins for these gene fragments. Subsequent studies support the involvement of chromatin structure; they show that, in nuclei, sites in non-transcribed genes recognized by these two transcription factors are less accessible to restriction enzyme digestion than the recognition sites in transcribed genes[15].

The differences between *in vitro* and *in vivo* binding of transcription factors to DNA that have been found in both yeast and *Drosophila* provide a framework for thinking about how the effective specificities of eukaryotic transcription factors arise. But the data also raise a host of questions. How many different processes influence the pattern of DNA binding *in vivo*? Are there general mechanisms and rules for all transcription factors, or is the binding of each class of factor affected by different mechanisms? Experiments that address specificity on a system-wide basis are in their infancy. Much work lies ahead for those interested in a global understanding of molecular interactions. □

1. Klotz, I.M. *Ligand Receptor Energetics: A Guide for the Perplexed* (John Wiley, New York, 1997).
2. Gilmour, D.S. & Lis J.T. *Mol. Cell. Biol.* **5,** 2009–2018 (1985).
3. Solomon, M.J., Larsen, P.L. & Varshavsky, A. *Cell* **53**, 937–947 (1988).
4. O'Brien, T., Wilkins, R.C., Giardina, C. & Lis, J.T. *Genes Dev.* **9**, 1098–1110 (1995).
5. Hecht, A., Strahl-Bolsinger, S. & Grunstein, M. *Nature* **383**, 92–96 (1996).
6. Strutt, H., Cavalli, G. & Paro, R. *EMBO J.* **16**, 3621–3632 (1997).
7. Wathelet, M.G. *et al. Mol. Cell* **1**, 507–518 (1998).
8. Carr, A. & Biggin, M.D. *EMBO J.* **18**, 1598–1608 (1999).
9. Ren, B. *et al. Science* **290**, 2306–2309 (2000).
10. Iyer, V.R. *et al. Nature* **409**, 533–538 (2001).
11. Lieb, J.D., Liu, X., Botstein, D. & Brown, P.O. *Nature Genet.* **28**, 327–334 (2001).
12. Johnson, A. in *Transcriptional Regulation* (McKnight, S.L. & Yamamoto, K.R., eds.) 975–1006 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1992).
13. Burns, L.G. & Peterson, C.L. *Mol. Cell. Biol.* **17**, 4811–4819 (1997).
14. Beato, M. & Eisfeld, K. *Nucleic Acids Res.* **25**, 3559–3563 (1997).
15. Carr, A. & Biggin, M.D. *Nucleic Acids Res.* **28**, 2839–2846 (2000).